Estimation of Missing Rainfall Data Using Spatial Interpolation and Imputation Methods

Noor Fadhilah Ahmad Radi[a], Roslinazairimah Zakaria[b] and Muhammad Az-zuhri Azman[c]

[a] *Institute of Engineering Mathematics, Universiti Malaysia Perlis, Taman Bukit Kubu Jaya, Jalan Seraw, 02000 Kuala Perlis, Perlis, Malaysia*
[b, c] *Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia*

Abstract. This study is aimed to estimate missing rainfall data by dividing the analysis into three different percentages namely 5%, 10% and 20% in order to represent various cases of missing data. In practice, spatial interpolation methods are chosen at the first place to estimate missing data. These methods include normal ratio (NR), arithmetic average (AA), coefficient of correlation (CC) and inverse distance (ID) weighting methods. The methods consider the distance between the target and the neighbouring stations as well as the correlations between them. Alternative method for solving missing data is an imputation method. Imputation is a process of replacing missing data with substituted values. A once-common method of imputation is single-imputation method, which allows parameter estimation. However, the single imputation method ignored the estimation of variability which leads to the underestimation of standard errors and confidence intervals. To overcome underestimation problem, multiple imputations method is used, where each missing value is estimated with a distribution of imputations that reflect the uncertainty about the missing data. In this study, comparison of spatial interpolation methods and multiple imputations method are presented to estimate missing rainfall data. The performance of the estimation methods used are assessed using the similarity index (S-index), mean absolute error (MAE) and coefficient of correlation (*R*).

Keywords: missing data; spatial interpolation; multiple imputations; rainfall data**.**
PACS: 92.40.eg

INTRODUCTION

Rainfall dataset may contain missing values which are due to malfunctions of instruments, bad weather or human error during data entry. Thus, an appropriate statistical method is highly in demand to solve missing data problem. Rainfall data plays a significant role in climatology as well as in agriculture, and important as a climatic parameter. Therefore, studies on rainfall data are significant in most countries due to lack of continuous data. It has been shown that studies in many countries revealed that the rainfall analysis is very crucial as a major factor influencing flood formation. Studies using rainfall data are really important for researchers and hydrologists in order to identify the characteristics of rainfall, the occurrence of spatial and temporal variability as well as the statistical modeling of predicting the occurrence of extreme rainfall events which leads to resolving problems such as floods, droughts and landslides. However, studies based on rainfall data could be disturbed by the occurrence of missing data. Therefore, studies of procedures and methods used to estimate missing data are essential.

The most common spatial interpolation method used in estimating missing rainfall data is normal ratio (NR) weighting method (Chow et al, 1988). This method is proposed by Paulhus and Kohler (1952) which is based on the past observations of the target station and neighbouring stations. A simpler method of spatial interpolation used in estimating missing data is arithmetic average (AA) weighting method. This method considers the average annual rainfall amount at the target station as well as the neighbouring stations and able to be used if the average annual rainfall amount at the target station is within 10% of the difference of the average annual rainfall amount from the neighbouring stations (Chow et al., 1988). The inverse distance (ID) weighting method is another simpler method which is based on the assumption that the rainfall amount at the target station could be influenced most by the nearest stations and less by the more distant stations (Suhaila et. al., 2008). The weight of ID method is made by finding the inverse distance of the neighbouring station to some power of the distances from the target station. In

most cases, power of 2 is used to obtain the weighted value. Teegavarapu and Chandramouli (2005) proposed the correlation coefficient (CC) weighting method by replacing the weighting value of the ID method with the correlation coefficient. Hence, the weighting value of ID method is $1/d_{it}^{b}$ whereas for CC is r_{it}. The study shows that the CC method is far more above superior than the traditional ID in interpolating the missing rainfall data. Alternative method for solving missing data is an imputation method. Imputation is a process of replacing missing data with substituted values. Instead of filling in a single value for each missing value, Rubin (2009) used multiple imputations procedure to replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These imputed data sets are combined and find an average. Then, the average of the total imputed data is used to fill in the missing data.

The objective of this study is to estimate missing data using both spatial interpolation and multiple imputation methods. In particular, we would like to estimate missing data by dividing the analysis into three different percentages namely 5%, 10% and 20% in order to represent various cases of missing data. Then, the performance of both methods are compared and assessed using the similarity index (S-index), mean absolute error (MAE) and coefficient of correlation (*R*).

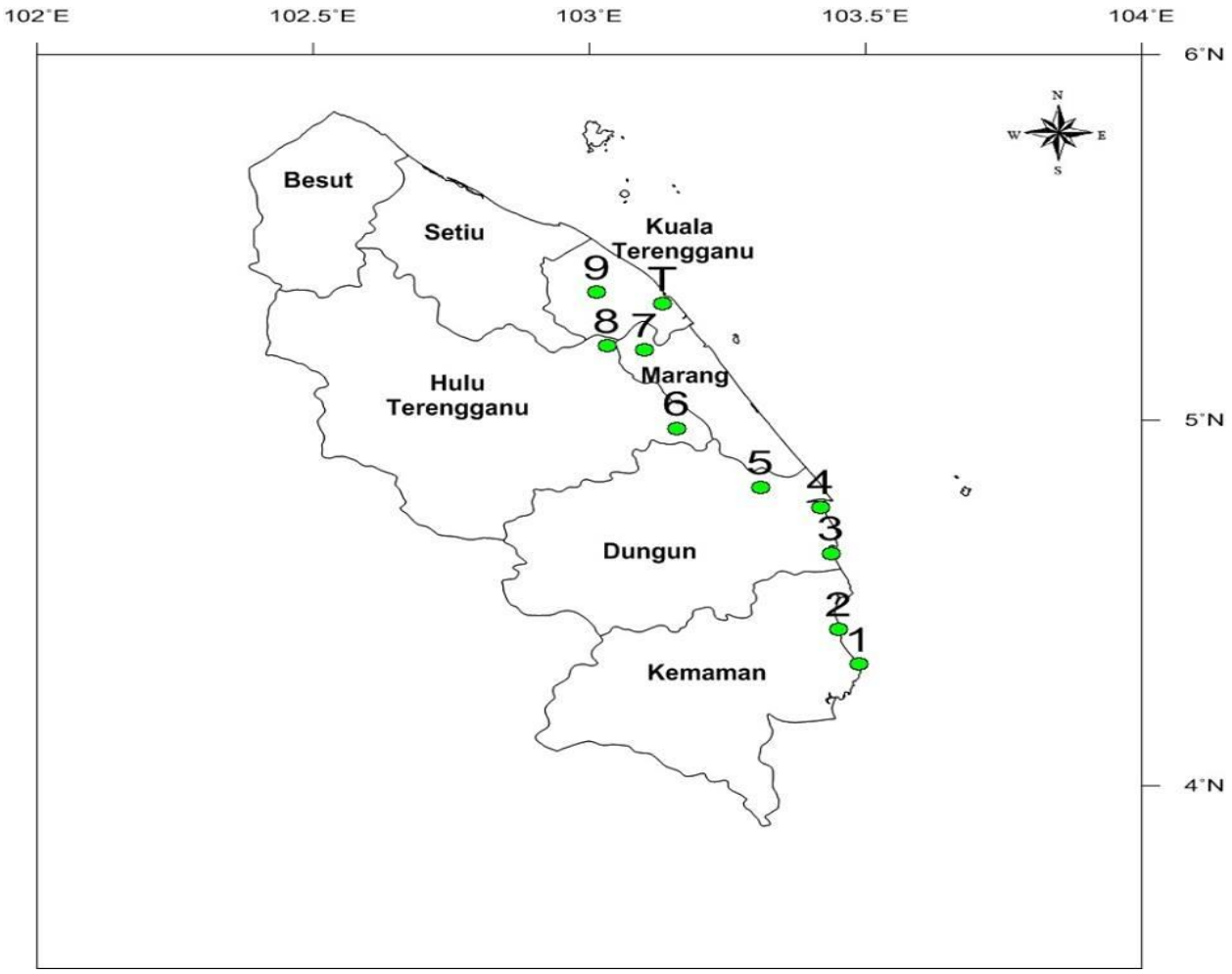

FIGURE 1. The location of the target station (T) and the neighbouring stations for rain gauges stations in Terengganu.

Area of Study

Kuala Terengganu which is the target station is located in the state of Terengganu. It is the capital city of Terengganu and located about 500 kilometers northeast of Kuala Lumpur and is bordered in the northwest by Kelantan, the southwest by Pahang and the east by the South China Sea. By having a tropical rainforest climate, Kuala Terengganu does not experienced a true dry season period, which classified this state as having a tropical monsoon climate. The average temperature is 26.7^0C with total average annual rainfall amount is 2911mm. Kuala Terengganu experiences dry season from May until June and February is the driest month with average annual rainfall of 60mm. Meanwhile, the rainy season occurs in November and December with average annual rainfall reaches more than 1000mm.

In this study, ten rain gauge stations in Terengganu are considered with Kuala Terengganu as the target station, refer to **FIGURE 2**. The data consist of daily rainfall amount from 1970 to 2012 (43 years). The data is obtained from the Department of Irrigation and Drainage, Ampang, Kuala Lumpur. Those years are chosen based on the completeness and longest available period of the rainfall data. The details of the location and correlation between particular target station and its corresponding neighboring stations are displayed in **TABLE (1)**.

TABLE (1). Description of the 10 rain gauge stations in the Terengganu within the radius 150km used as neighboring stations in this study with the target station in bold.

Station Number	Station Name	Latitude (^0S)	Longitude (^0E)	Euclidean Distance (km)	Correlation
T	**Setor JPS Kuala Terengganu**	**5.32**	**103.133**		
1	Sek.Keb. Kijal	4.33	103.488	1.0479 (116)	0.32
2	Sek.Men. Keb. Badrul Alam Shah	4.43	103.451	0.9455 (105)	0.38
3	Pusat Kesihatan Paka	4.64	103.438	0.7468 (83)	0.35
4	Sek. Men. Sultan Omar (Dungun)	4.76	103.419	0.6251 (69)	0.50
5	Rumah Pam Delong, Dungun	4.82	103.310	0.53162 (59)	0.38
6	Ldg. Koko Jerangau	4.98	103.158	0.3412 (20)	0.36
7	Sek. Men. Bkt. Sawa	5.19	103.100	0.1306 (15)	0.38
8	Sek. Keb. Kuala Telemong	5.20	103.032	0.1533 (17)	0.39
9	Sek.Keb. Kg. Gemuroh	5.35	103.014	0.1233 (14)	0.41

Research Methodology

In this section, we will briefly discuss methods for estimating missing data and assessing the performance of the methods used. The analysis involved a target and some selected neighbouring stations. In general, the target station has a complete set of data. Firstly, the missing data are identified in the neighbouring stations. For the neighbouring station that has missing data, the average value between the available data from the neighbouring stations is used. Then, missing data are introduced in the target station. Using the spatial and imputation methods, the missing rainfall data in target station are estimated and compared with the actual observations. The spatial interpolation and multiple imputation methods are as follows.

Spatial Interpolation and Multiple Imputations methods

(i) Arithmetic Average Method

The arithmetic average (AA) method is simply the average of rainfall amount of all the neighbouring stations. The estimated missing value is given by

$$p_t = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \tag{1}$$

where p_t is the estimated value of the missing rainfall at the target station, x_i is the observed rainfall at neighbouring station and n is the number of neighbouring stations.

(ii) Normal Ratio Method

This method is preferred if the average (or normal) annual rainfall of the station under consideration differs from the average annual rainfall at the neighbouring stations by more than 10%. The missing rainfall at the target station is estimated as the weighted average of neighbouring stations. The rainfall data at each of the neighbouring stations is weighted by the ratio of the average annual rainfall at the target station and average annual rainfall of the neighbouring station. The estimated missing value is given by

$$p_t = \frac{1}{n}\sum_{i=1}^{n}\frac{N_t}{N_i}x_i \; ; \tag{2}$$

where N_t is the annual rainfall amount at the target station and N_i is the annual rainfall amount at the *i*th neighbouring station.

(iii) Inverse Distance Method

This method weights neighbouring stations on the basis of their distance from the target station, on the assumption that closer stations are better correlated than those farther away. The estimated missing value is given by

$$p_t = \frac{\sum_{i=1}^{n} x_i / d_{it}^b}{\sum_{i=1}^{n} 1/d_{it}^b} ; \tag{3}$$

where d_{it} is the distance between the target station and the *i*th neighbouring station and b is the power of distance.

(iv) Coefficient of Correlation Method

This method is used by replacing the distance with the correlation coefficient as the weighting value. The estimated missing value is given by

$$p_t = \frac{\sum_{i=1}^{n} x_i \; r_{it}}{\sum_{i=1}^{n} r_{it}} ; \tag{4}$$

where r_{it} is the correlation coefficient of daily time series data between the target station and the *i*th neighbouring station.

(v) Multiple Imputation Method

A once-common method of imputation is single-imputation method which allows parameter estimation. However, the single imputation method ignored the estimation of variability which leads to the underestimation of standard errors and confidence intervals. To overcome underestimation problem, multiple imputations method is used, where each missing value is estimated with a distribution of imputations that reflect the uncertainty about the missing data. Multiple imputations provide one of the best methods in dealing with missing values. The same procedures of introducing missing data in the target station used in the spatial interpolation methods are applied in the multiple imputations method. The Amelia II package in R language version 3.0.2 is used to generate the imputed data sets. The Amelia package is based on bootstrap method. Since the rainfall data is always heavily skewed to the right, the data need to be transformed by taking the natural logarithm of the observed data before the method is applied. Then, the average of the imputed data set is calculated and used to fill in the missing data in the target station. In many studies, five imputed data sets are considered enough. For example, Schafer and Olsen (1998) suggest that in many applications, three to five imputations are sufficient to obtain excellent results. AmeliaView is an automated package for estimating missing data using multiple imputations but with limited options.

Performance of the estimation methods

In this study, the performance of the estimation methods used are compared and assessed using the similarity index (S-index), mean absolute error (MAE) and correlation coefficient (R). The error measures the difference between the estimation values and their corresponding observed values. The similarity index (S-index) is the index of agreement for assessing model performance which implies the percentage of agreement between the observed and estimated values. The values of S-index range from 0.0 for complete disagreement to 1.0 for perfect agreement (Wilmott, 1981). The three error indices are given by

$$\text{S-index} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{x}_i - x_i\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{x}_i - \bar{x}\right| + \left|x_i - \bar{x}\right|\right)^2} \tag{5}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{x}_i - x_i\right| \tag{6}$$

$$R = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(\hat{x}_i - \bar{x}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(\hat{x}_i - \bar{x}\right)^2}} \tag{7}$$

where x_i is the observed rainfall at neighbouring station, $\hat{x}_i$ is the estimated value and n is the number of neighbouring station.

Results and Discussion

In this section, we will briefly discuss the results of the analysis. The results of the performance of the estimation methods are shown in **TABLE (2)**. **FIGURE 2** shows graphical assessment of spatial interpolation and multiple imputation methods for various percentages of missing values using S-index, mean of absolute error (MAE) and correlation (R) methods. A good optimal distance should results in high values of S-index and correlation coefficient with low values of MAE.

TABLE (2). Comparison of estimation methods based on S-index, MAE and R with three different percentages of missing values.

Methods	**S-Index**			**MAE**			**Correlation Coefficient**		
	5%	10%	20%	5%	10%	20%	5%	10%	20%
NR	0.9931	0.9757	0.9710	0.3691	0.7471	1.3570	0.9864	0.9539	0.9450
AA	0.9929	0.9764	0.9703	0.3931	0.8004	1.5376	0.9858	0.9548	0.9427
ID	0.9923	0.9737	0.9640	0.4107	0.8143	1.6447	0.9848	0.9493	0.9301
CC	0.9930	0.9768	0.9709	0.3871	0.7913	1.5207	0.9861	0.9556	0.9438
MI	0.9932	0.9716	0.9650	0.3597	0.8179	1.4665	0.9865	0.9461	0.9333

In general, the result of the estimation method either for spatial interpolation and multiple imputation methods varies only at the third decimal place for both S-index and correlation coefficient values. The MAE results also consistent with the results obtained for S-index and MAE at various percentage of missing values. Three different percentages of missing data had been chosen, namely 5%, 10% and 20% respectively. As the percentages of missing data increased, the performance of each estimation methods tends to decrease for S-index and correlation coefficient resulting in the increment in MAE values. The NR method is found to be the best for both estimation methods used and the multiple imputations method is the second best based on their values of the three error indices. The lowest

performance is given by the ID method which is based on the distance between the target station and neighbouring stations. However, the ID method is considered good since it is not significantly difference from the other methods.

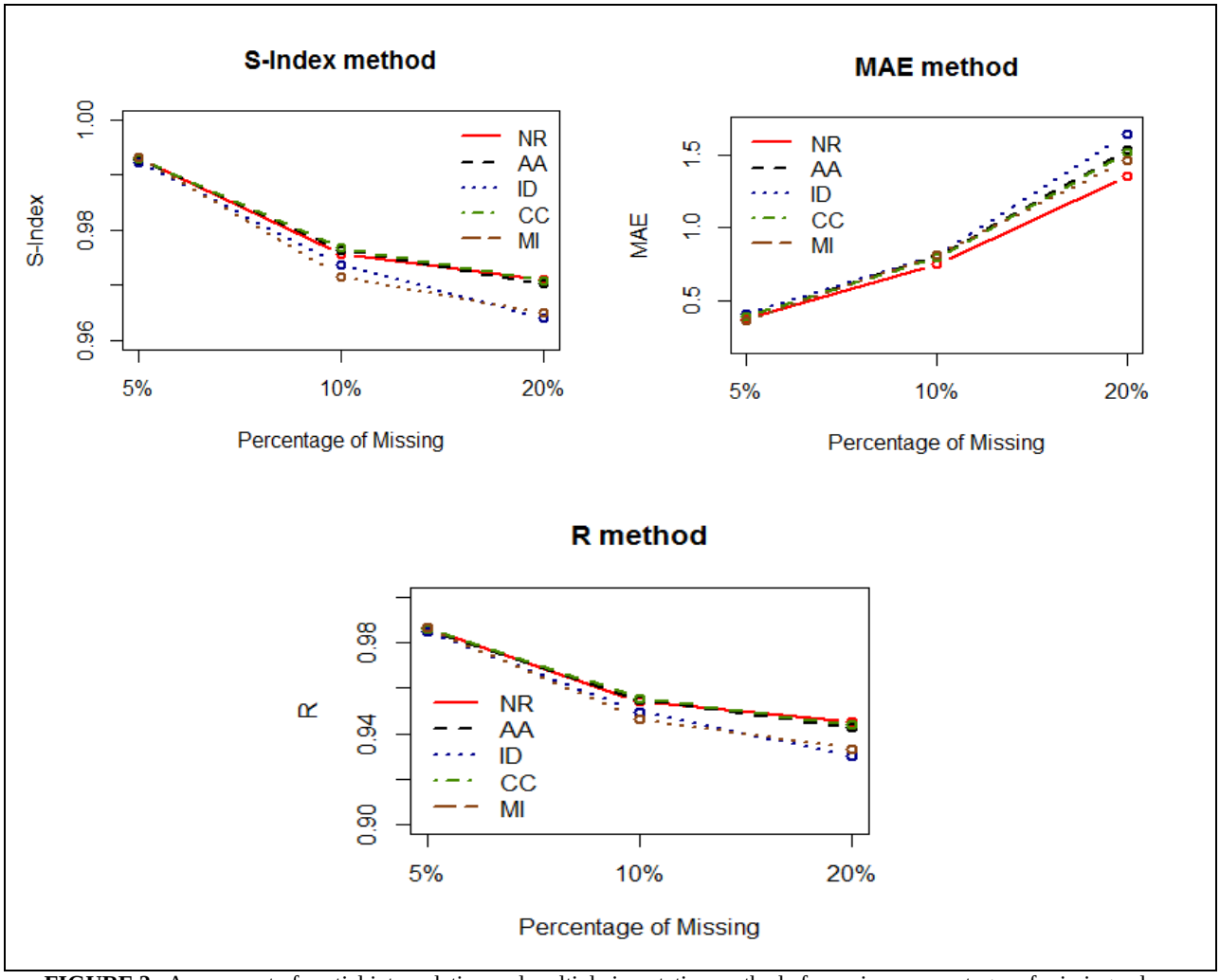

FIGURE 2. Assessment of spatial interpolation and multiple imputation methods for various percentages of missing values using S-index, mean of absolute error (MAE) and correlation (R) methods

Conclusion

In this study, the comparison of the spatial interpolation and the multiple imputations method are of the main interest. The spatial interpolation method includes normal ratio (NR), arithmetic average (AA), coefficient of correlation (CC) and inverse distance (ID) methods have been explored and applied using daily rainfall data from 1970 to 2012 for ten stations in Terengganu. The results are compared with the other estimation method that is multiple imputations (MI) method. All of these methods have been tested at three different percentages of missing data namely 5%, 10% and 20%, respectively. The results show that there are slightly increase in the value of mean absolute error (MAE) for each estimation method and decrease in values of S-index and correlation coefficient. Using various percentages of missing data, the results of the analysis do not have much effect. The NR and MI are found to be the best estimation methods among all to estimate missing rainfall data. For future study, one needs to consider other estimation methods such as mean imputation, regression-based method, resemblance-based "hot-deck imputation", expectation maximization (EM) and maximum likelihood methods. Other suggestions are increase the number of neighbouring stations involved as well as varying the distance between the target and neighbouring stations for estimating missing rainfall data.

ACKNOWLEDGMENTS

This study was supported by Universiti Malaysia Pahang (RDU120101).We thanks two anonymous referees whose comments led to a clearer presentation.

REFERENCES

1. C. J. Willmott, *Physical geography*, ***2*****(2)**, 184-194 (1981).
2. D. B. Rubin, *Multiple imputation for nonresponse in surveys* (Vol. 307), John Wiley & Sons, (2009)
3. J. L. Paulhus and M. A. Kohler, *Mon. Wea. Rev*, ***80***, 129-133 (1952).
4. J. Suhaila, S. M. Deni and A. A. Jemain, *Asia-Pacific Journal of Atmospheric Sciences*, ***44*****(2)**, 93-104 (2008).
5. J. L. Schafer, and M. K. Olsen, *Multivariate behavioral research*, ***33*****(4)**, 545-571 (1998).
6. R. S. Teegavarapu and V. Chandramouli, *Journal of Hydrology*, ***312*****(1)**, 191-206 (2005).
7. V.T. Chow, D.R. Maidment and L.W. Mays, Applied Hydrology, Mc Graw Hill Book Company, (1998).
8. Y. Xia, P. Fabian, A. Stohl, and M. Winterhalter, *Agricultural and forest meteorology*, ***96*****(1)**, 131-144 (1999).